Atti Accademia Nazionale dei Lincei

Classe Scienze Fisiche Matematiche Naturali

Rendiconti

Alfonso M. Liquori, Stefano Ottani, Alberto Ripamonti, Claudia Sadun

Genes as quasi-periodic linear lattices: a novel approach to the analysis of nucleotide sequences

Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti, Serie 8, Vol. **74** (1983), n.6, p. 389–395.

Accademia Nazionale dei Lincei

<http://www.bdim.eu/item?id=RLINA_1983_8_74_6_389_0>

Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti, Accademia Nazionale dei Lincei, 1983.

SEZIONE II

(Fisica, chimica, geologia, paleontologia e mineralogia)

Chimica. — *Genes as quasi-periodic linear lattices: a novel approach to the analysis of nucleotide sequences.* Nota di ALFONSO M. LIQUORI (*), STEFANO OTTANI (**), ALBERTO RIPAMONTI (**) e CLAUDIA SADUN (***), presentata (****) dal Socio V. CAGLIOTI.

RIASSUNTO. — Viene presentato un nuovo approccio allo studio di sequenze nucleotidiche basato su calcoli di serie Fourier.

La sequenza nucleotidica di un gene, rappresentabile mediante i simboli delle quattro basi (A, T, G, C) viene risolta in quattro sequenze « omonucleotidiche » che vengono considerate ciascuna come un reticolo lineare caratterizzato dalla ripetizione di una base e di « vacanze » (X) corrispondenti alle tre rimanenti basi. Viene calcolata la Trasformata di Fourier (F. T.). Successivamente vengono calcolate delle serie di Fourier aventi come coefficienti i quadrati dei valori assoluti della Trasformata di Fourier. Si ottiene così una funzione D (x) del tutto analoga alla « funzione Patterson », largamente impiegata nell'analisi strutturale dei cristalli.

I massimi della funzione D (x) corrispondono a vettori i cui moduli rappresentano distanze fra basi identiche nella sequenza nucleotidica del gene. Questa funzione viene mostrata per il gene della globina dell'*Aplisia*. La funzione contiene picchi corrispondenti a distanze di 3 unità e multiple di 3 (3, 6, 12, 15, ...).

One of the basic postulates of molecular biology states that a gene corresponds to a specific sequence (or primary structure) of a polynucleotide chain encoding the aminoacid sequence (or primary structure) of a polypeptide chain. The correspondence between nucleotide triplets of the gene and aminoacid residue of the polypeptide chain is defined by the genetic code. A gene may be interrupted by non coding intervening nucleotide sequences. However the transcription of the entire nucleotide sequence, followed by a splicing process, leads to a " mature " messenger RNA whose ribonucleotide sequence is complementary to the nucleotide sequence of the gene [1].

It should follow that the primary structure of a gene, either continuous or interrupted, should not display any sort of periodicity. On the other hand, a few reports may be found in the literature claiming that a periodicity (of every third base) has been detected within the primary structure of genes

(*) II Università di Roma, Torvergata. Centro Interdisciplinare dell'Accademia Nazionale dei Lincei. Via della Lungara 10, Roma.

(**) Istituto Chimico « G. Ciamician » Università di Bologna. Via Selmi 2, 40126 Bologna.

(***) Dipartimento di Chimica, Università « la Sapienza ». Piazzale Aldo Moro 5, 00185 Roma.

(****) Nella seduta del 14 maggio 1983.

and various explanations have been attempted in terms of a simpler primitive genetic code [2, 3]. This and other kinds of periodicities are however difficult to assess in consideration of the fact that there are only four bases and therefore there is always a rather high probability that periodic patterns may also appear in a purely statistic sequence.

We have therefore decided to explore the possibilities of quasi-periodic patterns in the primary structure of genes or small genomes by a rather general most suitable method based on Fourier vector analysis.

The primary structure of a gene resolved into four linear lattices

The nucleotide sequence (or primary structure) of a gene may be resolved (or decomposed) into four distinct linear lattices characterized by a unit spacing. They will correspond to a " resolved gene ". Each linear lattice corresponds to a homogeneous set of a given base (A , T, G or C). Within each linear lattice the three missing bases are represented by " lattices vacancies ". For instance the initial coding nucleotide sequence of the gene of *Aplysia* globin, corresponding to the ribonucleotide sequence of the mature RNA, may be represented by the following sets:

GCGGAGCCGGAGATCCGTAAAGGTCTTC ···
1) XXXXXXXXXXXXXTXXXTXXXXXTXTTX ···
2) XXXXAXXXXXAXAXXXXXAAAXXXXXX ···
3) XCXXXXCCXXXXXXCCXXXXXXXXCXXC ···
4) GXGGXGXXGGXGXXXXGXXXXGGXXXXX ····.

Clearly, whereas the symbols A , T , G ,C stand for the four bases, the symbol X stands for a vacancy (corresponding to the missing bases within each homonucleotide sequence). It should furthemore be noticed that the symbol T replaces U.

The Fourier Transforms of the four linear lattices corresponding to defective homonucleotide sequences

The above defined linear lattices may be studied by Fourier analysis, which has been largely employed for crystal lattices [4]. This approach requires the simulation of a linear " super lattice " characterized by a unit cell having a length L sufficiently larger than that of a linear lattice corresponding to the homonucleotide sequence diluited with " lattice vacancies ".

The Fourier transform of each linear lattice may by calculated as:

$$T(h) = \sum_j \delta_j \exp\left(2\pi i h \frac{x_j}{L}\right) \qquad j = 1, 2, \cdots, N. \tag{1}$$

Where h is an integer, L is the length of the unit cell of the super-lattice, x_j is the coordinate of the j^{th} base (or the vacancy) in the four homologous sequences, and $\delta_j = \begin{cases} 1 \\ 0 \end{cases}$ according to whether it refers to a base or to a vacancy (X) respectively in the linear lattice. The sum 1) is extended to the N lattice points (whose number corresponds to the number of nucleotides in the original gene).

A typical Fourier transform calculated for the "resolved gene" of *Aplisia* globin is shown in Fig. 1. As may be seen, a very rich pattern is obtained.

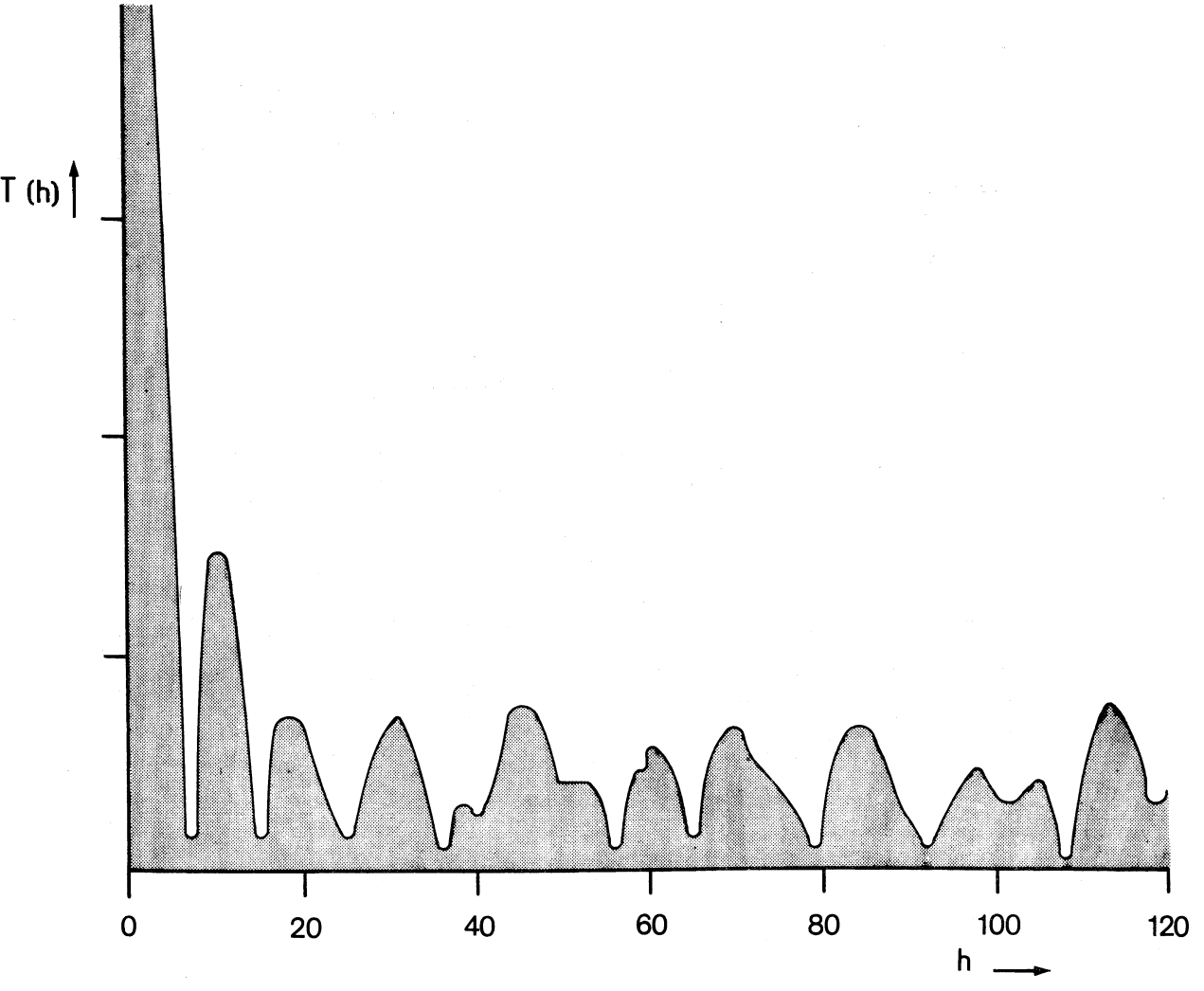

Fig. 1. – Modulus | T(h) | of the Fourier transform of a resolved gene of *Aplisia* globin consisting of a linear lattice containing at lattice points A or X (corresponding to the three missing bases T, G, C). The scale is arbitrary.

The autocorrelation function

The above calculated Fourier transform clearly indicates the presence of repeating patterns within the linear lattices consisting of separate sets of the bases with interdispersed vacancies.

In order to visualize such repeating patterns in a real unidimensional space, the autocorrelation functions were calculated as a Fourier series:

$$D(x) = \frac{2}{L} \sum_{h} |T(h)|^2 \cos\left(2\pi \frac{hx}{L}\right) \qquad h = 0, 1, \cdots \tag{2}$$

where x varies between 0 and L, $|T(h)|^2$ is the squared absolute value interpolated at a given value of h. The sum [2] is extended to a sufficiently large number of terms. In order to ensure convergency of the Fourier series, the $T(h)$ values were multiplied by a damping factor having the form $\exp - Bh^2$. The $D(x)$ function is strictly analogous to the Patterson Functions which have been largely employed in the structural analysis of crystal lattices by X-ray diffraction. The coefficients $|T(h)|^2$ are strictly analogous to the X-ray diffraction intensities (which are proportional to the absolute squared diffraction amplitudes). The physical meaning of a peak in the Patterson Function is that of the end of a vector whose origin lies at the origin of the unit cell and corresponds to an interatomic distance within the structure. Correspondingly, the meaning of a peak in the $D(x)$ function is that of the end of a vector between two identical bases contained in the above defined defective homonucleotide linear lattice.

The noise problem

Due to the finite length of a gene and to the limited number of different bases, any random polynucleotide sequences resolved into four homonucleotide lattices must be expected to yield $D(x)$ functions containing many peaks of variable heights.

In order to be able to attach any meaning to the $D(x)$'s corresponding to a real gene it is therefore necessary to consider the corresponding $D(x)$'s calculated for a random nucleotide sequence. The latter was generated according to current methods [5].

Results

Fig. 2 shows a plot of $D(x)$ function calculated for the gene of *Aplisia* globin (for $A = 1$) at " high resolution ". It contains a remarkable set of peaks which corresponds to a vector distance between every third base and

to its multiplies (3 – 6 – 9 – 12 – 15 – 18 ···). A plot of D (*x*) function (A=1) calculated for a random nucleotide sequence having the same base composition as that of the gene of the *Aplysia* globin is shown in Fig. 3.

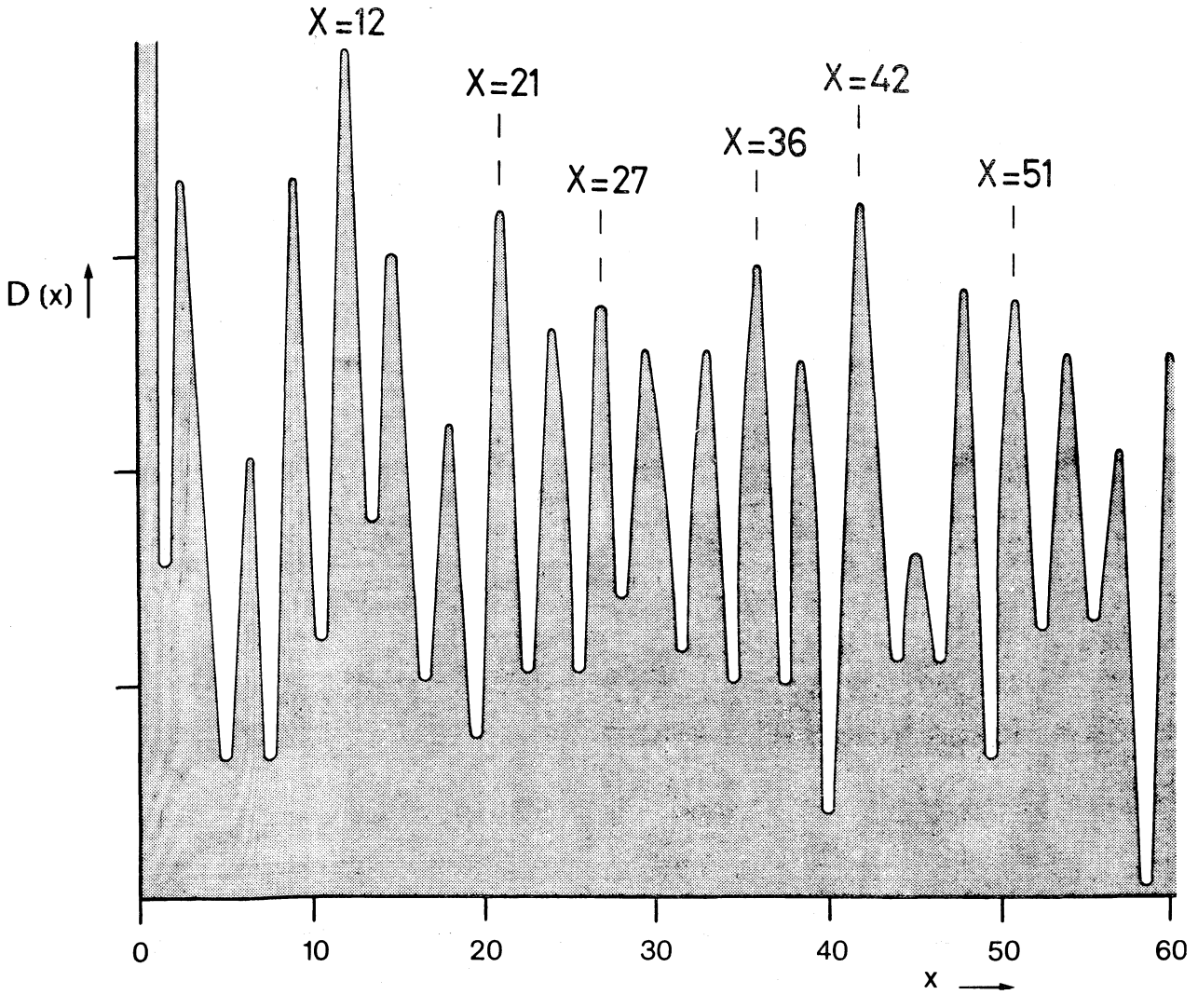

Fig. 2. – (D (*x*) function calculated at a " high resolution " for a resolved gene defined in Fig. 1.

As expected, the D (*x*) function of the random nucleotide sequence contains several peaks corresponding to various inter-base vectors. However the heights of the peaks are considerably lower than those of the gene and their position does not show any sharp multiplicity.

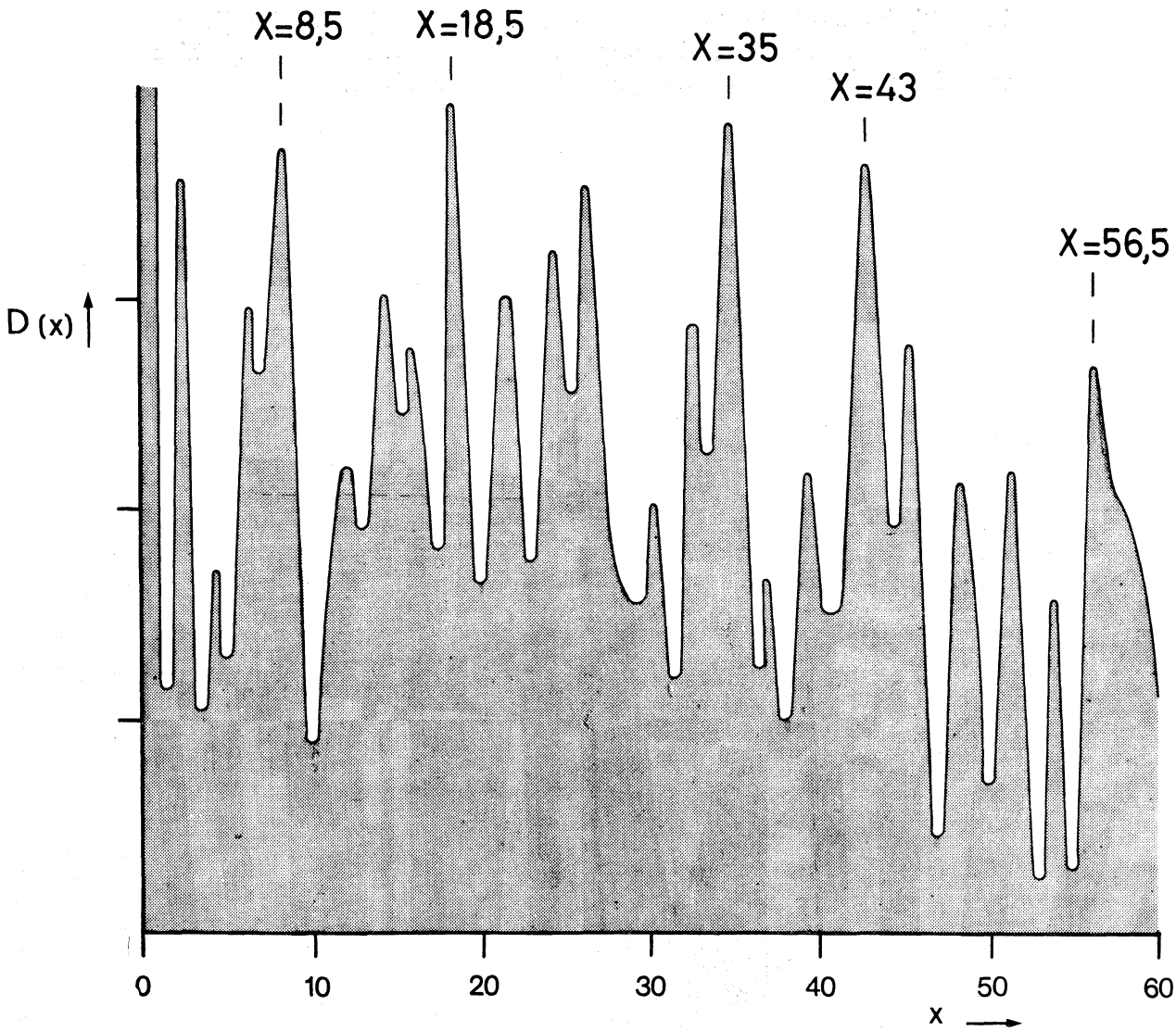

Fig. 3. - D (x) function calculated at the same resolution as in Fig. 2 for a linear random sequence of A and X.

A more complete account dealing with several genes and entire genomes will be given elsewhere.

This work has been carried out with the financial support of C.N.R., Italy.

REFERENCES

[1] B. ALBERTS, D. BRAY, J. LEWIS, M. RAFF, K. ROBERTS and J. D. WATSON (1983) – « Molecular Biology of the Cell », ed. Garland Publishing Inc. New York.
[2] J. C. W. SHEPHERD (1981) – « J. Mol. Evol. », *17*, 94.
[3] M. EIGEN and R. WINKLER (1981) – « Oswatish Die Naturwissenschaften », *68*, 217.
[4] A. M. LIQUORI (1960) – *La diffrazione dei raggi X nello studio della costituzione molecolare di sostanze naturali.* Varenna 1958 – Ed. Accad. Naz. Lincei.
[5] J. H. AHRENS, V. DIETER and A. GRUBE (1970) – « Computing », *6*, 121.